# Classification of message spreading in a heterogeneous social network

Siwar Jendoubi[1,2] ⋆, Arnaud Martin[2],
Ludovic Liétard[2], and Boutheina Ben Yaghlane[1]

[1] LARODEC, University of Tunis, Avenue de la liberté, 2000 Le Bardo, Tunisie
[2] IRISA, University of Rennes 1, Rue E. Branly, 22300 Lannion, France
jendoubi.siwar@yahoo.fr, Arnaud.Martin@univ-rennes1.fr,
ludovic.lietard@univ-rennes1.fr, boutheina.yaghlane@ihec.rnu.tn

**Abstract.** Nowadays, social networks such as Twitter, Facebook and LinkedIn become increasingly popular. In fact, they introduced new habits, new ways of communication and they collect every day several information that have different sources. Most existing research works focus on the analysis of homogeneous social networks, *i.e.* we have a single type of node and link in the network. However, in the real world, social networks offer several types of nodes and links. Hence, with a view to preserve as much information as possible, it is important to consider social networks as heterogeneous and uncertain. The goal of our paper is to classify the social message based on its spreading in the network and the theory of belief functions. The proposed classifier interprets the spread of messages on the network, crossed paths and types of links. We tested our classifier on a real word network that we collected from Twitter, and our experiments show the performance of our belief classifier.

**Keywords:** Information propagation, heterogeneous social network, classification, evidence theory

## 1 Introduction

Nowadays, social networks such as Twitter, Facebook and LinkedIn become increasingly popular. In fact, they introduced new habits and new ways of communication. Besides, one of the distinguishing features of on-line social networks is the information spreading through social links. This is due to the "word-of mouth" exchanges, *i.e.* user-to-user exchanges, which makes the information more accessible and it spreads and reaches a large scale in few minutes. The volume and the dynamic of the exchange has attracted the attention of research communities. This research is motivated by the fact that the study of the diffusion of information is useful for understanding the dynamic behind social networks and the evolution of human relationships. Thus, they have focused on

⋆ These research works and innovation are carried out within the framework of the device MOBIDOC financed by the European Union under the PASRI program and administrated by the ANPR.

the processing of such data to extract high quality information, this information may be an important event, it can also be useful for optimizing business performance, or even for preventing terrorist attacks, etc.

The processing of a social network, always, starts by studying its structural properties, in fact the simple visualization of the network cannot give us a clear analysis about it. In the literature, we found a lot of structural properties measures like the degree, the betweenness, the closeness, the eigenvector centrality, etc. Quantifying structural properties and interpreting them will be essential to characterize the behavior of social actors, their position in the network, their interactions and how do they diffuse the information. Hence, the analysis of the network structural properties is an essential step when we study and model information propagation.

In our work we are interested in the classification of the spreading of the information in a heterogeneous social network. We assume that each type of content has some specific behavior when it propagates in the network. Hence, we propose a new algorithm of information propagation in a heterogeneous social network that takes into account the behavior of the content to be propagated. Therefore, we introduce an evidential algorithm to classify the propagation of the information through the network.

In the next section, we outline the literature review of the information propagation in social networks, the social message classification and the theory of belief functions. In section three, we introduce our algorithm of information propagation in a heterogeneous social network. In section four, we present our classification algorithm. Finally, we present our experiments in the fifth section.

## 2 Literature review

### 2.1 Information propagation in social networks

Information dissemination is a wide research domain that attracted the attention of researchers from various field such as physics and biology. We find the family of epidemiological models that are used to understand how diseases spread through populations. The simplest version is SI (*Suspected-Infected*), in this model, an individual is suspected if he has not the disease yet but he can catch it and become infected. This model was extended and many other version appeared to model specific diseases. Hence, we find SIS model (*Suspected-Infected-Suspected*), SIR model (*Suspected-Infected-Recovered*), SIRS model (*Suspected-Infected-Recovered-Suspected*), etc. The reader can refer to [1, 17] for further details.

Computer scientists are generally interested in studying information propagation in on-line social networks. Mainly, their goal is to develop a model that simulates the diffusion process. Basic models are *Linear Threshold Model* (LTM) [7] and *Independent Cascade Model* (ICM) [6]. They assume the existence of a structure of a directed graph where each node can be activated or not knowing that you can not inactivate already activated nodes. The ICM model requires a probability distribution which must be associated with each link and LTM requires a degree of influence that must be set on each link and a threshold of influence for each node [12]. These two models were reused and improved in a lot of works like [5, 18].

In this paper, we focus on information propagation in a heterogeneous social network, *i.e.* on which we find several types of links and/or nodes. In fact, in real word social networks we find many types of objects (users, groups, applications, etc) that are connected *via* many types of social links (friendship, membership, colleague, etc). Information dissemination in homogeneous social networks has been widely studied and the reader can refer to [8] for a recent survey. Now, research works start focusing on the processing of heterogeneous social networks. We find the work of [19] that simulates the propagation of the information in heterogeneous social networks based on the configuration model approach. In [13], authors propose to consider the behavior of individuals to model the influence propagation, their model is based on a heterogeneous social network.

### 2.2 Social message classification

Social message classification approaches, presented in the literature, are generally based on the content of the information and text mining techniques. They search to classify the user generated content to positive or negative about a some specific product. This task is so called sentiment classification and it is used to mine opinions. It starts by an item and/or feature extraction step, then it compares the extracted items and/or features to an existing corpus, finally comes the sentiment classification that can be based on items, features or both of them [14]. We find the work of [15] in which the author used a random sample of 3516 tweets to classify the feelings of consumers with respect to well-known brands. He classified the opinions (tweets) into positive and negative to see what is the most dominant opinion. In [10], a detailed case study that applies text mining to analyze unstructured textual content published on Twitter and Facebook and that talks about three chains of pizza. The reader can refer to the work of [16] for a recent study of the state of the art of social networks data mining.

### 2.3 Theory of belief functions

*Upper and Lower probabilities* [4] was the first ancestor of the theory of belief functions. Then comes the *Mathematical theory of evidence* [20] which defines the basic framework of information management and processing in the evidence theory, often called *Shafer model.* The main purpose of the theory of belief functions is to achieve more reliable, precise and coherent information. Here we present a short introduction of this theory, for more details the reader can refer to [20].

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a set of all possible decisions that can be made in a particular problem, it is called frame of discernment. The basic belief assignment (BBA), $m^\Omega$, represents the agent belief on $\Omega$, and it must respect $\sum_{A\subseteq\Omega} m^\Omega(A) = 1$. In the case where we have $m^\Omega(A) > 0$, $A$ is called focal set

**Algorithm 1** Information propagation algorithm

**Inputs:**
- **N:** number of iteration
- **S:** source of the message
- **Str:** propagation strategy
- **Network:** the heterogeneous social network

**Output:**
- **PrNet:** propagation network

**Algorithm:**
1. ReadyNodes.add($S$);
2. **For** $i = 1$ **to** $N$ **do**
   (a) **for** $j = 1$ **to** ReadyNodes.size() **do**
      i. Node ←ReadyNodes.get($j$);
      ii. if(Node.propagate()=True)
         **foreach** LinkType **do**
            $x \leftarrow$ Node.outdegree() $*$ Node.propagationTendancy()
               $*$Str.LinkTypeProportion();
            R← (Node.randomSelection($x, LinkType$));
   (b) Pr.refine($R$);
   (c) R1.addAll($R$);
   (d) ReadyNodes.addAll($R1$);
   (e) R1.clear;

of $m^{\Omega}$. The basic belief assignment can be converted into other functions defined from $2^{\Omega}$ to $[0, 1]$. This theory presents a rich framework for information fusion and combining pieces of information (evidence). We find the Dempster's rule [4], the conjunctive and disjunctive combination rule [21], etc.

# 3 Propagation algorithm in a heterogeneous social network

In this section we introduce an algorithm of information propagation in a heterogeneous social network. This new algorithm takes four different inputs which are the number of iterations (stopping condition), the source of the message, the propagation strategy and the heterogeneous social network. As output, we have the propagation network that preserves the traversed paths. Algorithm 1 shows outlines of our propagation process. It starts by the source node. First, we verify if the current node is ready (wants) to propagate the message. Then, for each type of link in the network we compute the number of neighbors that will receive the message.

We assume that each type of message has some special characteristics of propagation in the network that is related to the types of links, so we define a propagation strategy for each type of message. Moreover, we consider the tendency of a particular node to propagate the message as a propagation parameter. Indeed, this parameter models the fact that a node can choose to distribute the message to a subset of its contacts (that he selects) or to retain it. The novelty of this algorithm is that we consider the type of the message while propagating it. Moreover our algorithm works with heterogeneous social networks where we have different types of links.

## 4 Classification of information propagation

The main purpose of this paper is to classify the spreading of the information through the network in order to characterize its content. In this section, we introduce our classification process that is composed of two steps; parameter learning step and the classification step. As mentioned in the algorithm 2, to learn the parameters of the model we need a set of propagation networks. First of all, we compute the number of nodes that have received the message *via* each type of link. We do this computation for each propagation level, *i.e.* we call propagation level the number of links between the source of the message and the target node. Second, we calculate the accrued effective by summing the effective of each level with the effective of the one before, this computation is done in order to preserve the propagation history at each propagation level. After that we transform the effective set of each level to a probability distribution defined on types of links, this transformation is done for two reasons; the first one, we need a probability distribution for the probabilistic classifier and the second one, it is an essential step to get the basic belief assignment distribution. Finally we transform each probability distribution to a BBA distribution using the consonant transformation [2, 3].

**Algorithm 2** Parameter learning algorithm

**Input:**
- **PrNetSet:** a set of propagation networks

**Output:**
- **ProbaSet:** a set of probabilities distributions (a probability distribution by propagation level).
- **BbaSet:** a set of BBA distributions (a BBA distribution by propagation level).

**Algorithm:**

//effective computation
**Foreach** PrNet **in** PrNetSet **do**
1. **Foreach** Level **in** PrNet **do**
   (a) **Foreach** TypeLink **do**
   $N$ (TypeLink, Level) $\leftarrow N$ (TypeLink, Level) +ComputeNodes($TypeLink$);

//Accrued effective calculation
**For** Level= 2 **to** NbrLevels **do**
1. **Foreach** TypeLink **do**
   (a) $N$ (TypeLink, Level) $\leftarrow N$ (TypeLink, Level) $+N$ (TypeLink, Level $-$ 1);

//ProbaSet and BbaSet computation
ProbaSet$\leftarrow$ProbabilitiesCalculation($N$);
BbaSet$\leftarrow$ConsonantTransformation(ProbaSet);

Once model's parameters are learned, we can use it to classify new coming message (propagation network of the message) as shown in algorithm 3. Our classification algorithm starts by applying the same parameter learning process (algorithm 2) on the propagation network to be classified. Then for each level in the network we compute the distance between its probability distribution and the probability distribution of each propagation strategy, then we choose the class of the nearest propagation strategy (with the shortest distance) to be the class of the message in the current level. The same process is done with BBA distributions as mentioned in the algorithm.

**Algorithm 3** Classification algorithm

**Input:**
- **ProbaSets:** a set of probabilities distributions for each strategy of propagation.
- **BbaSets:** a set of BBA distributions for each strategy of propagation.
- **PrNet:** The propagation network to be classified

**Output:**
- In order to see the impact of the level of propagation on the classification results, in our output we have a class by level.

**Algorithm:**
1. (ProbaPr, BbaPr) $\leftarrow$ ParameterLearning ($PrNet$);
2. **For** $i = 1$ **to** NbrStrategies **do**
   (a) **Foreach** Level **do**
      i. ProbaDist($i, Level$) $\leftarrow$ Distance (ProbaPr, ProbaSets ($i$));
      ii. BbaDist($i, Level$) $\leftarrow$ Distance (BbaPr, BbaSets ($i$));
3. **Foreach** Level **do**
   (a) ProbaClasses($Level$) $\leftarrow$ StrategyMinDistance (ProbaDist (:, $Level$));
   (b) BbaClasses($Level$) $\leftarrow$ StrategyMinDistance (BbaDist (:, $Level$));

## 5 Experiments and results

In this section, we present some experiments to show the power of the proposed evidential classification algorithm.

### 5.1 Data description

We used NodeXL V 1.0.1.245 [9] to collect social network data from Twitter. We collected the network shown in figure 1. It is a directed network in which nodes are Twitter users. Table 1 shows the characteristics of our network data.

As mentioned above, we need a heterogeneous social network to test proposed algorithms. Therefore, we used the structure of the network collected from Twitter and we generated, randomly, the types of links. We assume four types of link in the network which are “Professional”, “Familial”, “Friendly” and “Undefined”. Then we obtained a heterogeneous social network that is used as input for our propagation algorithm.

Table 1: Data characteristics

| Vertices | Edges | Geodesics distance | Betweenness | Closeness | Eigenvector |
|---|---|---|---|---|---|
| 97 | 350 | 6 | 184.99 | 0.004 | 0.01 |

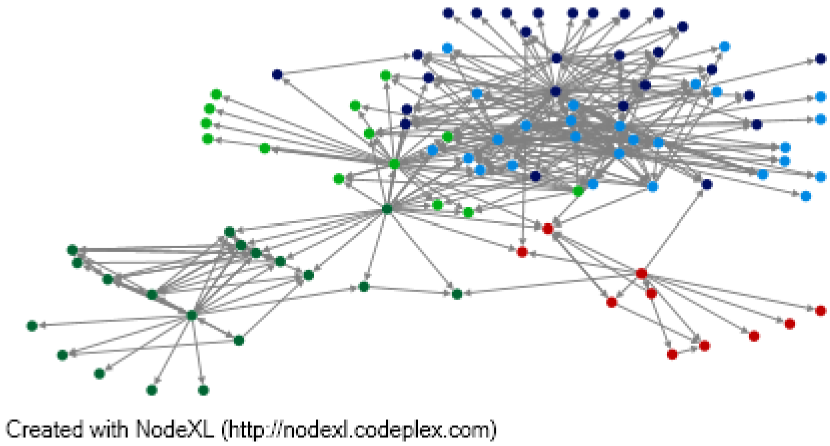



Fig. 1: Network visualization

### 5.2 Experiment configuration

In the following experiments, we defined three different propagation strategies for three types of messages which are: "Spam", "Professional" and "Familial". Each strategy is defined as the proportion of the nodes that will receive the message from each type of links. Hence, we have to define four proportions for each propagation strategy. To be as near as possible to the reality, we added a noise rate to the strategy. We note that the noise value can be added or removed from the proportions of kind of messages. We used the euclidean distance for the probabilistic classifier:

$$d_E\left(Pr_1, Pr_2\right) = \sqrt{\sum_{i=1}^{card}\left(Pr_1\left(i\right) - Pr_2\left(i\right)\right)^2} \tag{1}$$

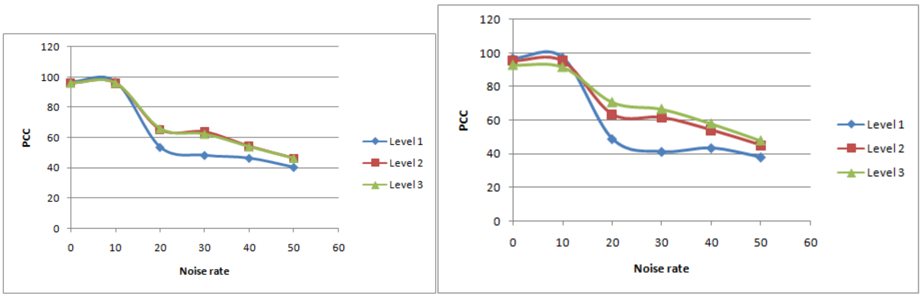
and the Jousselme distance [11] for the evidential one:

$$d_J\left(m_1, m_2\right) = \sqrt{\frac{1}{2}\left(m_1 - m_2\right)^T \underset{=}{D}\left(m_1 - m_2\right)} \tag{2}$$

such that $\underset{=}{D}$ is an $2^n \times 2^n$ matrix and $D\left(A, B\right) = \frac{|A \cap B|}{A \cup B}$. We fixed the number of levels in the network to three (three iterations in the propagation algorithm). Then we run the proposed propagation algorithm to create a training set for each propagation strategy, we fixed the size of the strategy training set to 100 propagation networks. Also, we created a testing set of size 100.

### 5.3 Results and discussion

In this section we present our results and a comparison between the probabilistic and the evidential classifier. To obtain accurate results we turned the experimental process ten times and we take the mean of the percentage of correctly classified (PCC) propagation networks. Figure 2 shows the impact of the propagation level on the PCC of the probabilistic results (figure 2a) and the evidential results (figure 2b). Figures 2a and 2b illustrate that the PCC increases when the propagation level increases and we observe this fact starting from the noise level 20%. In figure 2a we observe that the curve of the second level coincides with the curve of third level and practically there is no improvement in the PCC. However, in figure 2b (evidential results), we note that the PCC increases with the propagation level, this fact is observed starting from the noise rate 20%. Hence, we have the PCC of the third level greater than the PCC of the first and the second levels, and the PCC of the second level is higher than the PCC of the first one. Therefore, more the message propagates in the network, more we can characterize it.




(a) Probabilistic results (b) Evidential results

Fig. 2: The impact of the propagation level on the PCC

In figure 3, we compare the probabilistic and the evidential results of the third propagation level. We note that without noise (0%) the probabilistic PCC is about 96% (with a 95% confidence interval of $\pm 1.27$) and the evidential PCC is equal to 93% (with a 95% confidence interval of $\pm 1.60$), but in real world social networks the absence of the noise is an ideal fact and cannot be realistic. When the noise rate increases, the curve shows that the percentage of correctly classified propagation networks (messages) decreases. However, we see that the evidential (Belief) PCC starts to be greater than the probabilistic (Proba) one. We observe this fact from the noise rate 20% where we have an evidential PCC equals to 70.7% (±4.33) and a probabilistic PCC equals to 65.8% (±4.18). Thus, we can conclude that the evidential classifier is more robust against the noise and gives better classification rates than the probabilistic classifier.

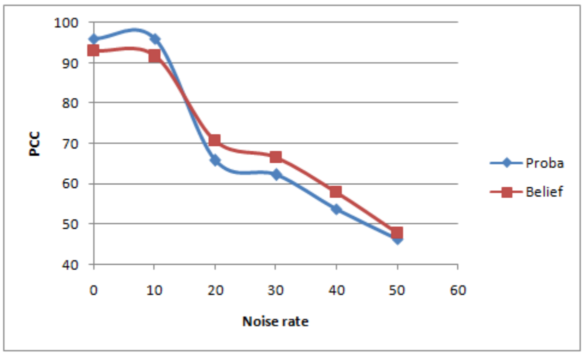



Fig. 3: Comparison between probabilistic results and evidential results (level three)

## 6 Conclusion

To conclude, we presented a state of the art of the information propagation, classification of social messages and the evidence theory. Then, we proposed an algorithm of information propagation in a heterogeneous social network. Thereafter we introduced a new evidential classification approach that classifies message propagation in a heterogeneous social network. Finally, we presented some experiments and we noticed the performance of the evidential classifier against the probabilistic one in noisy cases. Moreover, we observed that when the propagation level increases, the message class becomes more accurate and more realistic.

For future works, we will compare our propagation algorithm with previous algorithms. Also, we will search to improve it by the management of the uncertainty and the imprecision related to types of relationships between social actors. Our next goal is therefore to define a message propagation algorithm that takes into account the uncertainty of the types of relationships that is defined on the links, also we will search to consider the heterogeneity of nodes in the network. Second, we will run our classification algorithm with a more complex heterogeneous social network in order to prove its applicability.

## References

1. Anderson, R.M., May, R.M.: Infectious Diseases of Humans. Oxford University Press (1991)

2. Aregui, A., Denoeux, T.: Consonant belief function induced by a confidence set of pignistic probabilities. In: ECSQARU. pp. 344–355 (October 2007)
3. Aregui, A., Denoeux, T.: Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. Int. J. Approx. Reasoning 49(3), 575–594 (2008)
4. Dempster, A.P.: Upper and Lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W.: Outtweeting the twitterers - predicting information cascades in microblogs. In: WOSN'10. pp. 3–11 (2010)
6. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12(3), 211–223 (August 2001)
7. Granovetter, M.: Threshold models of collective behavior. American journal of sociology pp. 1420–1443 (1978)
8. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. SIGMOD Rec. 42(1), 17–28 (March 2013)
9. Hansen, D.L., Shneiderman, B., Smith, M.A.: Analysing social media network with nodeXL insights from a connected world. Elsevier Inc. (2011)
10. He, W., Zhab, S., Li, L.: Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management 33, 464–472 (2013)
11. Jousselme, A.L., Grenier, D., Bossé, E.: A new distance between two bodies of evidence. Information Fusion 2, 91–101 (2001)
12. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: KDD'03 Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 137–146. ACM Press (2003)
13. Li, C.T., Lin, S.D., Shan, M.K.: Influence propagation and maximization for heterogeneous social networks. In: WWW'12-Poster Presentation. pp. 559–560 (April 2012)
14. Lo, Y.W., Potdar, V.: A review of opinion mining and sentiment classification framework in social networks. In: Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference (Juin 2009)
15. Mostafa, M.M.: More than words: Social networks text mining for consumer brand sentiments. Expert Systems with Applications 40, 4241–4251 (2013)
16. Nettleton, D.F.: Survey data mining of social networks represented as graphs. computer sciences review 7, 1–34 (2013)
17. Newman, M.E.J.: Networks: An introduction. Oxford University Press (2010)
18. Saito, K., Ohara, K., Yamagishi, Y.and Kimura, M., Motoda, H.: Learning diffusion probability based on node attributes in social networks. In: ISMIS'11. pp. 153–162 (2011)
19. Sermpezis, P., Spyropoulos, T.: Information diffusion in heterogeneous networks: The configuration model approach. In: INFOCOM, 2013 Proceedings IEEE. pp. 3261–3266 (April 2013)
20. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
21. Smets, P.: Belief Functions: the Disjunctive Rule of Combination and the Generalized Bayesian Theorem. International Journal of Approximate Reasoning 9, 1–35 (1993)